Applied Factor Analysis in the Natural Sciences

RICHARD A. REYMENT
K. G. JÖRESKOG
University of Uppsala
Appendix by LESLIE F. MARCUS
American Museum of Natural History
Queens College, CUNY

CAMBRIDGE
UNIVERSITY PRESS

PUBLISHED BY THE PRESS SYNDICATE OF THE UNIVERSITY OF CAMBRIDGE
The Pitt Building, Trumpington Street, Cambridge CB2 IRP

CAMBRIDGE UNIVERSITY PRESS
The Edinburgh Building, Cambridge CB2 2RU, United Kingdom
40 West 20th Street, New York, NY 10011-4211, USA
10 Stamford Road, Oakleigh, Melbourne 3166, Australia

Original edition published in 1976 by Elsevier Scientific Publishing Company, Amsterdam
Second edition published by Cambridge University Press 1993

First published 1993
First paperback edition 1996

Printed in the United States of America

Library of Congress Cataloging-in-Publication Data is available.

A catalog record for this book is available from the British Library.

ISBN 0-521-41242-0 hardback
ISBN 0-521-57556-7 paperback

Contents

1 Introduction

1.1 STRUCTURE IN MULTIVARIATE DATA

Commonly, almost all natural scientists make a great number of measurements in their daily activities, for example, on the orientation of strata, geochemical determinations, mineral compositions, rock analyses, measurements on fossil specimens, properties of sediments, ecological factors, genetics, and many other kinds. You need only reflect on the routine work of a geological survey department in order that the truth of this statement may become apparent.

Scientific data are often multivariate. For example, in a rock analysis, determinations of several chemical elements are made on each rock specimen of a collection. You will all be familiar with the tables of chemical analyses that issue from studies in igneous petrology and analytical chemistry. Petrologists have devised many kinds of diagrams in their endeavor to identify significant groupings in these data lists. The triangular diagrams of petrology permit the relationships between three variables at a time to be displayed. Attempts at illustrating more highly multivariate relationships have led to the use of ratios of elements and plots on polygonal diagrams (cf. Aitchison, 1986).

Obviously, one can only go so far with the graphical analysis of a data table. The logical next step is to use some type of quantitative method for summarizing and analyzing the information hidden in a multivariate table. It is natural to enquire how the variables measured for a homogeneous sample are connected to each other and whether they occur in different combinations, deriving from various relationships in the population. One may, on the other hand, be interested in seeing how the specimens or objects of the sample itself are interrelated, with the thought in mind of looking for natural groupings. In both cases, we should be seeking structure in the data.

Geologists and biologists came into touch with the concept of factor analysis and the study of multivariate data structure through the contacts between paleontologists and biologists. The biologists, in their turn, learnt the techniques from psychometricians. Thus, the French zoologist Teissier studied multivariate relationships in the carapace of a species of crabs (Teissier, 1938), using a centroid first-factor solution of a correlation matrix. He interpreted this "general factor" as one indicating differential growth.

Let us now look briefly at a few typical problems that may be given meaningful solutions by an appropriately chosen variety of eigenanalysis.

A geochemist has analyzed several trace elements in samples of sediment from a certain area and wishes to study the relationships between these elements in the hope of being able to draw conclusions on the origin of the sediment.

A mining geologist is interested in prospecting an area for ores and wants to use accumulated information on the chemistry and structural geology of known deposits in the region to help predict the possibilities of finding new ore bodies.

A paleontologist wishes to analyze growth and shape variation in the shell of a species of brachiopods on which he has measured a large number of characters.

A petroleum company wants to reduce the voluminous accumulations of data deriving from paleoecological and sedimentological studies of subsurface samples to a form that can be used for exploring for oil-bearing environments.

In an oceanological study, it is desired to produce graphs that will show the relationships between bottom samples and measurements made upon them on a single diagram, as a means of relating organisms to their preferences for a particular kind of sediment.

1.2 AN EXAMPLE OF PRINCIPAL COMPONENT FACTOR ANALYSIS

At this point, we think it would be helpful to you if we gave you an inkling of what is obtained in a principal component factor analysis. (The reason for making this distinction will become clear later on.) We have chosen an artificial mining example by Klovan (1968) because it not only introduces the geological element at an early stage but also provides a good practically oriented introduction to the subject.

Imagine the following situation. We wish to carry out exploration for lead and zinc in an area containing a high-grade lead–zinc ore. The area has been well explored geologically and the bedrock is made up of an altered carbonate–shale sequence. The map area and the sampling grid are displayed in Fig. 1.1.

The three controls, paleotemperature (T), strength of deformation of the bedrock (D), and the permeability of the rock (P) are considered to determine the occurrence of lead and zinc, for the purposes of our example. It is assumed that these controls are determinable from

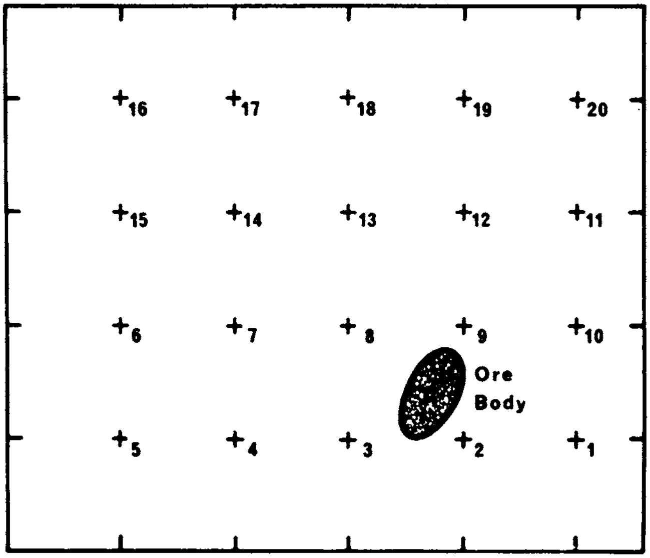

Figure 1.1. The sampling grid for the prospecting example.

observations on 10 chemical, mineralogical, and rock-deformational variables. The distribution of these causes will in reality never be known but, for this example, we shall imagine that they are distributed as shown in Fig. 1.2. You will note that the lode lies at the intersection of these causes at certain specified levels. These are, for paleotemperature, 80–90, for deformation, 35–45, and for permeability of the country rock, 45–50. Accepting that a geological survey of the area would have given as clear an indication as our manufactured example, it would not be unreasonable to expect that target areas for intensive prospecting would occur in localities where the intersection situation is repeated.

The three controls cannot, of course, be estimated directly. They can, however, be measured indirectly from geological properties that are a reflection of them. The arrays shown in Table 1.I list the artificial data, as well as the information used in constructing this set of observations. The left array of numbers gives the "amount" of each of the three controls at each of the localities; the upper array states precisely the degree to which each of the geological variables is related to the causes. Multiplication and summation of every row of the left array by every column of the top array yields the large array (corresponding to raw data) at the bottom. Naturally, in a real study, you would not know the left-hand and top arrays of Table 1.I. All you would have at your disposal would be the large array, or data matrix, the result of a detailed geological survey and a laboratory study of the samples collected.

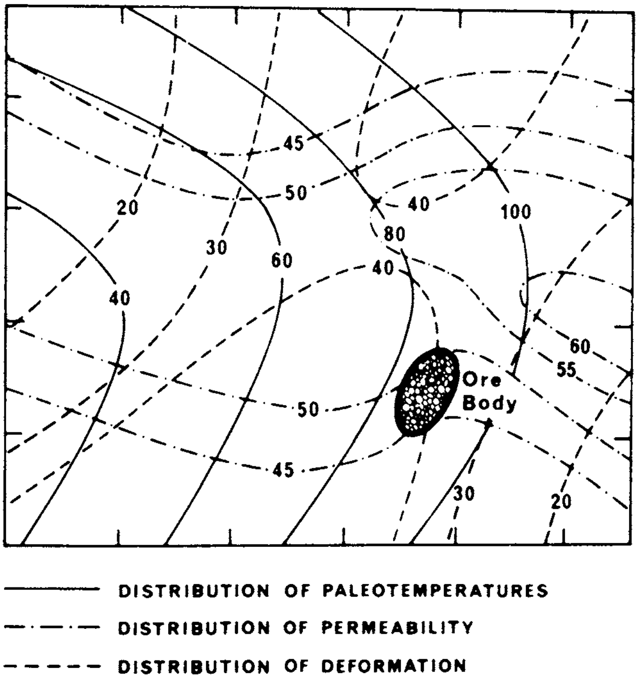

Figure 1.2. Distribution of controls imposed at the locations of the samples.

The question to be answered now is, how can we determine the existence of structure in such a large array of numbers? The technique of factor analysis turns out to be a useful way of providing plausible answers.

Simply put, factor analysis creates a minimum number of new variables, which are linear combinations of the original ones such that the new variables contain most or all of the information.

The starting point is provided by the correlations between the variables measured, 10 in all. The matrix of correlation coefficients is listed in Table 1.II. It was subjected to principal component factor analysis for which three significant factors were obtained. Thus, we began with 10 characters but can now "explain" the total variability of the sample in terms of 3 new variables or factors.

The principal-factor matrix is listed in Table 1.III; it shows the "composition" of the factors in relation to the original variables. As these factors are usually not readily interpretable, it is accustomed practice to rotate the reference axes by some appropriate method in order to bring out the important contributing loadings and to diminish the loadings on nonsignificantly contributing variables. The visual result

Table *1.1. The raw data matrix for the lead–zinc prospecting problem*

				Geological properties									
			Causes	Mg in calcite	Fe in sphalerite	Na in muscovite	Sulfide	Crystal size of carbonates	Spacing of cleavage	Elongation of ooliths	Tightness of folds	Vein material per m^2	Fractures per m^2
			T	0.95	0.75	0.75	0.33	−0.20	0.05	0.20	0.10	0.00	0.05
			D	0.00	0.10	0.20	0.33	0.60	0.95	0.70	0.85	0.10	0.25
			P	0.05	0.15	0.05	0.34	0.60	0.00	0.10	0.05	0.90	0.70
Locality	T	D	P	Data matrix									
1	121	21	46	117.25	99.75	97.25	62.50	16.00	26.00	43.50	32.25	43.50	43.50
2	96	35	42	93.30	81.80	81.10	57.51	27.00	38.05	47.90	41.45	41.30	42.95
3	78	54	49	76.55	71.25	71.75	60.22	46.20	55.20	58.30	56.15	49.50	51.70
4	63	51	49	62.30	59.70	59.90	54.28	47.40	51.60	53.20	52.10	49.20	50.20
5	42	44	44	42.10	42.50	42.50	43.34	44.40	43.90	43.60	43.80	44.00	43.90
6	39	26	54	39.75	39.95	37.15	39.81	40.20	26.65	31.40	28.70	51.20	46.25
7	52	36	52	52.00	50.40	48.80	46.72	42.40	36.80	40.80	38.40	50.40	48.00
8	67	46	54	66.35	62.95	62.15	55.65	46.60	47.05	51.00	48.50	53.20	52.65
9	90	37	51	88.05	78.85	77.45	59.25	34.80	39.65	49.00	43.00	49.60	49.45
10	108	27	61	105.65	92.85	89.45	65.29	31.20	31.05	46.60	36.80	57.60	54.85
11	112	33	59	109.35	96.15	93.55	67.91	32.80	36.95	51.40	42.20	56.40	55.15
12	91	38	59	89.40	80.90	78.80	62.63	40.00	40.65	50.70	44.35	56.90	55.35
13	76	39	54	74.90	69.00	67.50	56.31	40.60	40.85	47.90	43.45	52.50	51.35
14	63	30	51	62.40	57.90	55.80	48.03	36.00	31.65	38.70	34.35	48.90	46.35
15	43	19	55	43.60	42.40	38.80	39.16	35.80	20.20	27.40	23.20	51.40	45.40
16	68	16	42	66.70	58.90	56.30	42.00	21.20	18.60	29.00	22.50	39.40	36.80
17	77	27	41	75.20	66.60	65.20	48.26	25.40	29.50	38.40	32.70	39.60	39.30
18	93	37	43	90.50	79.90	79.30	57.52	29.40	39.80	48.80	42.90	42.40	44.00
19	102	47	48	99.30	88.40	88.30	65.49	36.60	49.75	58.10	52.55	47.90	50.45
20	120	36	46	116.30	100.50	99.50	67.12	25.20	40.20	53.80	44.90	45.00	47.20

Table *1.II. Correlations among the* 10 *geological properties*

	Mg	Fe	Na	S	Crystal	Cleavage	Ooliths	Folds	Veins	Fractures
1	1.000									
2	0.998	1.000								
3	0.994	0.998	1.000							
4	0.905	0.931	0.941	1.000						
5	−0.574	−0.519	−0.493	−0.171	1.000					
6	0.125	0.178	0.231	0.479	0.628	1.000				
7	0.572	0.618	0.657	0.833	0.274	0.883	1.000			
8	0.275	0.328	0.377	0.610	0.533	0.988	0.945	1.000		
9	0.011	0.056	0.035	0.288	0.541	0.196	0.228	0.222	1.000	
10	0.256	0.312	0.313	0.593	0.556	0.540	0.616	0.588	0.907	1.000

Table *1.III. Results of the factor analysis*

Variable	Communality	Factors 1	2	3
Principal factors of the correlation matrix				
1	1.0000	0.7933	−0.6029	0.0850
2	1.0000	0.8302	−0.5501	0.0902
3	1.0000	0.8505	−0.5247	0.0373
4	1.0000	0.9742	−0.2073	0.0892
5	1.0000	0.0376	0.9992	−0.0111
6	1.0000	0.6609	0.5989	−0.4522
7	1.0000	0.9310	0.2365	−0.2781
8	1.0000	0.7693	0.5001	−0.3976
9	1.0000	0.3389	0.5369	0.7726
10	1.0000	0.6765	0.5364	0.5046
Variance		54.614	31.928	13.459
Cumulative variance		54.614	86.542	100.000
Varimax factor matrix				
1	1.0000	0.9974	0.0716	0.0056
2	1.0000	0.9916	0.1201	0.0479
3	1.0000	0.9839	0.1771	0.0245
4	1.0000	0.8776	0.3992	0.2656
5	1.0000	−0.6206	0.5952	0.5105
6	1.0000	0.0533	0.9882	0.1435
7	1.0000	0.5125	0.8391	0.1823
8	1.0000	0.2055	0.9637	0.1705
9	1.0000	0.0014	0.0534	0.9986
10	1.0000	0.2227	0.4059	0.8864
Variance		44.66	33.41	22.00

of the rotation will then be that some of the loadings will have been augmented while others will have become greatly lower. In our example, we used the varimax rotation technique. The varimax factor matrix displayed in Table 1.III demonstrates what we have just described, and you will see this if you compare entries in the two upper listings of the table, entry by entry. The rotated factor matrix contains 10 rows and 3 columns, each latter representing a factor. Reading down a column, the individual numbers tell us the contribution of a particular variable to the composition of the factor; in fact, each column can be thought of as a factor equation in which each loading is the coefficient of the corresponding original variable.

Table *1.III* (*cont.*)

Locality	Factors 1	2	3
Varimax factor score matrix			
1	1.6820	−1.1085	−0.8445
2	0.5988	0.2442	−1.3125
3	−0.2144	1.8308	0.0837
4	−0.7787	1.5001	0.0481
5	−1.5367	0.8901	−0.8567
6	−1.5277	−1.0713	0.5476
7	−1.0955	−0.0325	0.3471
8	−0.5739	0.9167	0.8009
9	0.3768	0.2012	0.1881
10	1.1764	−0.9354	1.6917
11	1.2853	−0.3045	1.4412
12	0.4317	0.1081	1.5106
13	−0.1745	0.2776	0.7082
14	−0.6279	−0.5465	0.1032
15	−1.3205	−1.7494	0.6197
16	−0.3619	−1.6486	−1.5528
17	−0.0900	−0.5506	−1.5758
18	0.4709	0.4016	−1.1222
19	0.7671	1.2609	−0.1759
20	1.5394	0.3161	−0.6498

A third chart of numbers emerges from the factor analysis, the varimax factor score matrix, shown in Table 1.III. This gives the amounts of the new variables at each of the sample localities. With this matrix, we are able to map the distributions of these new factor variables on the sample grid.

It requires sound geological reasoning in order to interpret the results of a factor analysis. From Table 1.III, you will see that the first factor is mainly concerned with the variables "Mg in calcite," "Fe in sphalerite," and "Na in muscovite," a combination indicating temperature dependence. The second factor is heavily loaded with the variables "spacing of cleavage," "elongation of ooliths," and "tightness of folds," a combination speaking for rock deformation. The third factor is dominated by the variables "vein material/m^2" and "fractures/m^3," interpretable as being a measure of permeability of the country rock.

The distribution of the three sets of factor scores is shown in Fig. 1.3. The patterns of Fig. 1.2 are almost exactly duplicated. By comparing the nature of the intersections around the known ore body, and searching the diagram for a similar pattern, you will see that at least one other

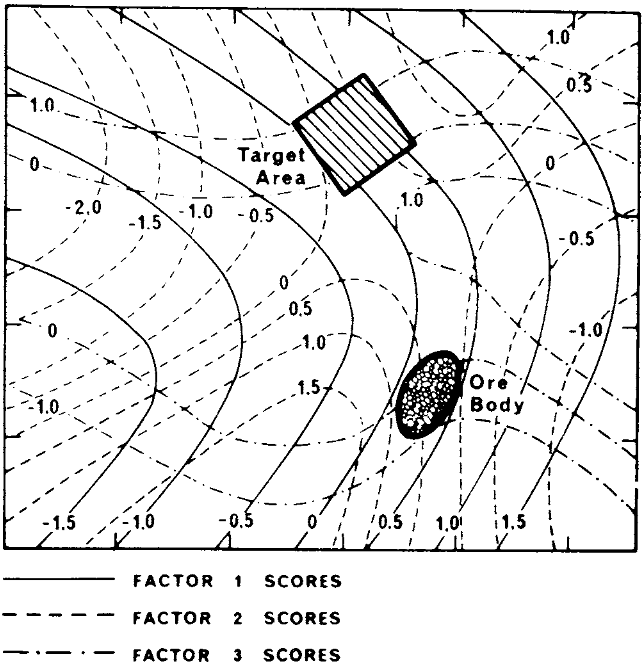

Figure 1.3. Map of the composite factor scores for the three factors of the ore-prospecting example.

area on the map has the same special conditions. The marked square is thus the first-order target for further exploration. This is an artificial example, contrived to give a good result. Under actual exploration conditions, you would not expect things to fall out so nicely and your geological knowledge would be put to a greater test. This is only a brief summary of the example. If you want to work through all the steps, we refer you to Klovan (1968).

You will find a good deal of simple algebra in the ensuing pages, much of which you might think unnecessary, bearing in mind the ubiquitousness of computer programs for doing most of the things occurring in factor analysis. Obviously, nobody today is going to suggest seriously that you try inverting matrices, or extracting eigenvalues and eigenvectors, on your own, although we have devoted some space to the arithmetic of this topic. There are numerous excellent programs for doing these calculations at any computer installation and a wide selection of PC software. It is our considered opinion, nevertheless, that you should have some idea of what is done by the computer in performing these operations. Moreover, despite the fact that many varieties of factor analysis are available at most installations, not all of them are to be recommended for general use. We have therefore made a point of introducing you to the most useful and mathematically best defined procedures so that you will be able to make a satisfactory choice among the programs for factor analysis available to you.

Several recent updatings of principal component analysis have appeared. It is relevant to our revision to see what topics have been taken up in those texts.

Jolliffe (1986) covered some of the fields mapped out in the First Edition. He included the highly appropriate data-analytical topics of robust estimation procedures, determination of the "correct" number of principal components, influential observations, and the isolation of atypicalities in the data. He correctly noted that although *true factor analysis* and *principal component analysis* may, in some respects, have similar aims, they are different techniques. This was clearly enunciated in the First Edition, but we opted for a nonspecific data-analytical use of the concept of "factor analysis," which permits greater latitude in discussing methods that reduce to an eigensolution.

One of the bones of contention in principal component analysis is the question of *rotating* eigenvectors in the manner usually thought proper to true factor analysis. Jolliffe (1986, p. 118) observed that there may be circumstances in which rotation of a subset of principal components can prove advantageous – the main positive effect of this maneuver is that it tends to simplify the factor loadings, or rotated principal component coefficients, without the implication that a factor model is being assumed. The most recent treatment of this subject is Chapter 8 in Jackson (1991).

The volume by Flury (1988) provides an uncontroversial introduction to principal component analysis. Such debatable questions as the rotation of eigenvectors are not discussed at all. The main theme of that text concerns *common* principal component analysis, a variant of standard principal components, in which attention is paid to differences in covariance matrices due to "pooling" (hence the adjective "common") and the distributional theory of eigenvalues and eigenvectors. Canonical correlation is included as a peripheral technique.

The book by Preisendorfer (1988), completed posthumously by C. D. Mobley, gives an interesting insight into applications in meteorology and oceanography. Canonical correlation is also included among the techniques considered. Rather surprisingly, Preisendorfer's way of treating some problems discloses ignorance of well-established methods of multivariate statistical analysis as well as a lack of familiarity with standard terminology. This deficiency occurs despite the fact that the appropriate references appear in his bibliography. An example is the problem of testing that an eigenvector agrees with a given vector. Factor analysis is also included but without exemplifying the role of true factor analysis (perhaps weather forecasting could yield a suitable example). Rotation of eigenvectors is accepted as a matter of course by Preisendorfer as is also Procrustean superposition of "vector frames."

The volume compiled by Fornell (1982) contains several valuable pointers to future areas of development. One of the illuminating topics is that concerning the revival of interest in the method of *path analysis* of Sewall Wright. Other subjects covered are canonical correlation and the analysis of redundancy. This topic is also taken up by Jöreskog and Sörbom (1989).

Another informative volume is that of Digby and Kempton (1987). This is an ecologically oriented treatment in the hands of statisticians. It encompasses many of the methods that we gather beneath the umbrella of factor analysis and that were included in the First Edition: these include methods of ordination, principal coordinates, correspondence analysis, and the analysis of asymmetry. A very recent reference is the text on principal component analysis by Jackson (1991). As a matter of interest, the rotation of axes to simple structure is not disputed as being a useful technique.

The slim volume by Gordon (1981) is a compact reference, replete with essential information for the methods of principal components, principal coordinates, correspondence analysis, and Gabriel's biplot. Useful sources of information on factor analysis in the geosciences are to be found in the journals *Mathematical Geology* and *Computers & Geosciences*. Articles of biological significance appear regularly in *Biometrics*, *Biometrika*, and *Evolution*. The applied scientist can find much of interest in the pages of *Applied Statistics*, in which problems of biological relevance appear frequently.

1.3 OVERVIEW OF PROBLEMS AMENABLE TO TREATMENT BY FACTOR-ANALYTICAL TECHNIQUES

The present section reviews briefly a randomly chosen set of articles in which factor analysis of some variety has been used in order to solve a scientific problem. We do this in the hope that this will give you a better insight into the types of problems amenable to treatment by the factor class of techniques.

Relationships between organisms and sedimentary facies

In a study of the Pleistocene–Holocene environment of the northwestern part of the Persian Gulf, Melguen (1971) used correspondence analysis to explore relationships between ecological and sedimentological facies in the estuaries of the Rud Hilla and Rud Mund. The study material was derived from sediment cores taken from depths ranging between 8 and 60 m. Thirty-three components were determined on samples from the cores, including counts on the abundances of shell-bearing organisms, serpulid tubes, fish remains, plants, fecal pellets, argillaceous lumps, minerals, and rock fragments.

Petrology

Saxena (1970) studied a multicomponent, multiphase system of minerals by using the principal components of the correlation matrix and plotting the transformed observations. On such representations, he demonstrated that by plotting certain relative positions of all coexisting minerals as well as the components of the multiphase system, lines joining points representing pairs of coexisting minerals are significant in the same sense as in concentration diagrams (Gibb's triangle, for instance). There is a clear advantage offered by the components approach in that the lines stand for the influence of all the components of the system.

Sedimentary petrology

Osborne (1967, 1969) has employed factor analysis for grouping Ordovician limestones, on the basis of characters determinable in thin sections. He succeeded (Osborne, 1967) in attaching paleoecological significance to the factors extracted. Within much the same frame of reference, McCammon (1968) made a comparative study of factor-analytical methods of grouping facies of Recent carbonates of the Bahama Bank. A similar study for Jurassic limestones of the northern Alps has been done by Fenninger (1970).

Mineralogy

Middleton (1964) used factor analysis to elucidate a complicated mineralogical problem in scapolites. By applying principal components and factor analysis to major- and trace-element data, he could identify the marialite–meionite solid solution in scapolites and propose the possible existence of an independent end-member bearing Mg and OH. Three significant groupings of trace elements were deduced. Mineralogical analyses often require special treatment (Aitchison, 1986).

Stratigraphy

R-mode factor analysis was used by McElroy and Kaesler (1965) on well data from the Upper Cambrian Reagan Sandstone on the Central Kansas Uplift. The four factors extracted were interpreted in terms of subsidence during the time of deposition of the sandstone, regional distributional patterns, and periods of uplift or nonsubsidence.

Biofacies relationships

Cheetham (1971) made *R*- and *Q*-mode factor analyses of weight-percentage abundances of major biotic constituents in a calcareous mound in the Danian of southern Scandinavia (transforming to minimize the constant sum constraint). From the *R*-mode loadings, three kinds of influences involving bryozoans and corals could be recognized, as well as the spatial relationships of biofacies.

Intertidal environment

A question of paleoecological significance concerns the identification of communities in tidal sediments. Cassie and Michael (1968) tried several multivariate methods in a well-documented study of this problem and came to the conclusion that principal component analysis proved to be the most versatile of them in that it permits both the diagnosis of the community structure and a plausible contouring of the communities in space.

Heavy minerals

Imbrie and Van Andel (1964) studied occurrences of heavy minerals from the Gulf of California and the Orinoco–Guyana shelf by *R*- and *Q*-mode factor analysis. The two areas have quite different sedimentological histories. Factor analysis of the simple situation represented by the Californian material yielded results similar to those obtained by conventional inspection, although more meaningful detail was revealed. The more remote petrographical sources of the Orinoco–Guyana shelf produce a more complicated situation with much mixing of the minerals. The factor analysis yielded a mineral distributional pattern greatly different from the impression given by mere inspection of the data. These results could be interpreted convincingly in terms of transportation during the post-Pleistocene rise of sea level.

Vertebrate paleontology

Gould (1967), analyzing pelycosaurs by *R*- and *Q*-mode factor analysis, was able to demonstrate far-reaching agreement for his computational results with the accepted scheme of phylogeny. Mahé (1974), in what is essentially a review of a comprehensive study of Madagascan fossil lemurs, advocated the pilot application of correspondence analysis to a few specimens of a sample in order to identify the most meaningful variables for a multivariate analysis. Using this variety of factor analysis, he established the phylogenetic relationships among lemurs, using craniometrical characters, but noted that the approach works best at the generic level.

Geochemistry of magmas

Teil and Cheminée (1975) analyzed major and trace elements in Ethiopian lavas by correspondence analysis, whereby meaningful associations between elements and samples could be shown to exist. The results turned out to be in agreement with accepted chemically based considerations for fractionation of magmas.

Palynology

In a study of Flandrian pollen data, Birks (1974) made a principal component analysis of percentage data on frequencies of pollen types. The component scores for the individual samples were plotted in relation to the stratigraphical position of the sample, thus forming composite "curves" of the original pollen variables. Birks makes here a highly significant suggestion, namely, that "a pollen zone can be delimited on the basis of stratigraphically adjacent samples with similar compositional scores."

Geochemistry of Cambrian alum shale

Armands (1972) studied in detail the geochemistry of uranium, molybdenum, and vanadium in Swedish alum shale. In this treatise, principal components and factor analysis were used to determine the paragenesis of elements in alum shales. Partly with the help of the results of these analyses, Armands found that Upper and Middle Cambrian alum shales can be divided into five categories, notably, detrital, authigenic, carbonate, sulfide, and organic.

Paleoecology

Variations in the relative frequencies of different species in samples may be interpreted in terms of major environmental factors to which the organisms react. Reyment (1963) used principal components and factor analysis to unravel paleoecological relationships between environmental forces and 17 species of Paleocene ostracods. The statistical analysis succeeded in separating euryoic species from stenoöic ones (see Section 8.3). Birks and Gordon (1985) give several examples of multivariate paleoecological studies.